Report File Glossary

Report Type

For Affymetrix expression chips, the entry in this field will be Expression Report. The example .RPT file given here is from an expression chip. Other analyses are also available in MAS 5.0; these yield different types of .RPT files. For more information on the other types of .RPT files, see the Affymetrix Microarray Suite 5.0 User's Guide.

Date

The date and time of your scan will be reported here.

Filename

The name of your sample will be reported here with a .CHP extension.

Probe Array Type

This is the type of Affymetrix chip you are using (e.g. HG-U133A).

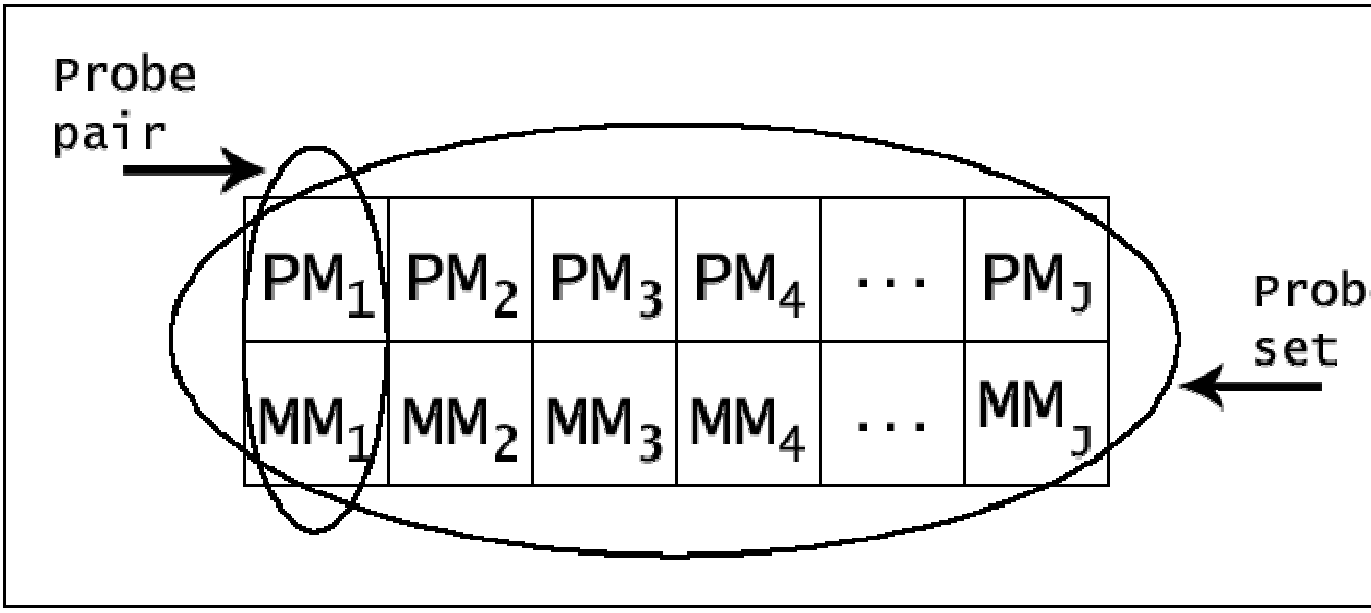
Algorithm

The Statistical Algorithms are the collection of algorithms used in Affymetrix Microarray Suite 5.0 to calculate Presence Calls (Present, Marginal, or Absent) and Signal for each gene represented on the chip. Previous versions of Affymetrix GeneChip Analysis Suite and Microarray Suite used the Empirical Algorithms. For more information on the Presence Calls, see the Presence Calls section of this website. For more information on the Signal values, see the Expression Estimates section of this website.

Probe Pair Thr

This is the minimum number of probe pairs (threshold) a probe set must have for the probe set data to be included in the calculation of the report statistics. A probe set refers to the entire collection of J PM/MM probe pairs that measure the expression of a particular mRNA transcript.

A probe pair is determined to be unusable if the MM pair is saturated, or if the PM and MM values are within τ of each other (see the Tau section for a description of τ . The number of usable probe pairs must meet or exceed the probe pair threshold for an expression estimate to be reported for the probe set.



Controls

Sense or Antisense. This refers to the direction of the probe sequences on the array. Most Affymetrix probes are antisense since labeling for antisense probes is easier than for sense probes. There are some sense arrays, for example the *E. coli* sense array.

Alpha1 and Alpha 2

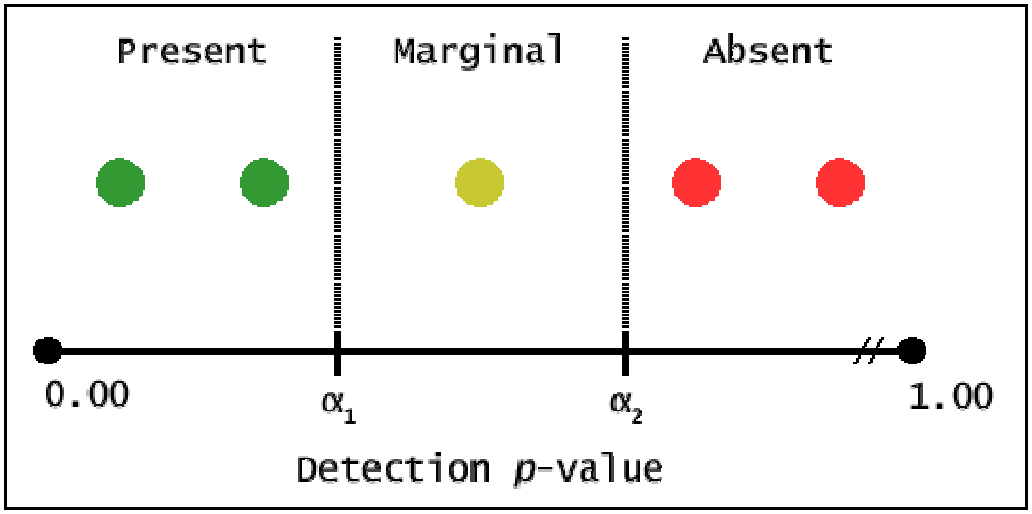
α_1 and α_2 are cutoffs for the Detection *p*-value that are used to determine Presence Calls of Present, Marginal, or Absent in an absolute expression analysis. In this example for the HGU-133A chip, the default values are α_1=0.05 and α_2=0.065. If Detection *p*-value < α_1, an mRNA transcript is determined to be Present in the sample. If α_1 < Detection *p*-value < α_2, the transcript is determined to be Marginal. If α_2 < Detection *p*-value, the transcript is determined to be Absent.

The Detection *p*-value is calculated as follows. For each of the *J* PM/MM pairs in a probe set, calculate a Discrimination score R_j:

$$R_j = (PM_j - MM_j)/(PM_j + MM_j) \text{ for } j=1,...,J.$$

Note that if the PM_j intensity is much larger than the MM_j intensity, R_j will be close to 1. On the other hand, if the PM_j and MM_j intensities are close to each other, R_j will be close to 0, and possibly negative. The *J* Discrimination scores are then compared to a user-defined threshold τ (by default τ=0.015) and are ranked according to their distance from τ. The One-Sided Wilcoxon's Signed Rank test is then used to test for a significant difference from τ, yielding a Detection *p*-value for the gene. Probe sets with many R_j values close to 1 will yield lower (more significant) Detection *p*-values, whereas those with many R_j values close to 0 will yield higher (less significant) Detection *p*-values.

Decreasing α_1 can reduce the number of false detected calls (increase specificity), but it may also reduce the number of true present calls (decrease sensitivity).



gene1 and gene2 are Present since ρ-value < α_1
gene3 is Marginal since α_1 < ρ-value < α_2
gene4 and gene 5 are Absent since ρ-value < α_2

Tau

τ is a parameter that is used in performing the One-Sided Wilcoxon's Signed Rank test for the Detection call. It represents a threshold that the Discrimination score for a probe set must exceed in order for a gene to be regarded as Present in the sample. For each of the *J* PM/MM pairs in a probe set, calculate a Discrimination score R_j:

$$R_j = (PM_j - MM_j)/(PM_j + MM_j) \text{ for } _j=1,...,J.$$

Note that if the PM_j intensity is much larger than the MM_j intensity, R_j will be close to 1. On the other hand, if the PM_j and MM_j intensities are close to each other, R_j will be close to 0, and possibly negative. The *J* Discrimination scores are then compared to a user-defined threshold τ (by default τ=0.015) and are ranked according to their distance from τ. The One-Sided Wilcoxon's Signed Rank test is then used to test for a significant difference from τ, yielding a Detection *p*-value for the gene. Probe sets with many R_j

values close to 1 will yield lower (more significant) Detection *p*-values, whereas those with many R_j values close to 0 will yield higher (less significant) Detection *p*-values.

Increasing τ may reduce the number of false positives (increase specificity), but it may also reduce the number of true present calls (decrease sensitivity).

Noise (RawQ)

Noise (RawQ) measures the degree of pixel-to-pixel variation of probe cells on a GeneChip array and is calculated by the Empirical Algorithms used in versions of MAS previous to MAS 5.0. It remains in the .RPT file to accommodate users who previously used Noise (RawQ) as a quality control measure. Noise, reported later in the .RPT file, is the Statistical Algorithm analogue to Noise (RawQ).

Noise (RawQ) is calculated using the pixel-level data in the .DAT file as follows:

$$Q = (1/N) * (\Sigma\ stdev_i/\sqrt{pixel_i}) * SF * NF.$$

- N = number of background cells
- $stdev_i$ = standard deviation of all pixels in the ith background cell
- $pixel_i$ = number of pixels in the ith background cell
- SF = scaling factor for the analysis
- NF = normalization factor for the analysis

Electrical noise of the scanner and sample quality can both contribute to noise. Arrays coming from the same scanner should ideally have comparable noise (RawQ) values.

Scale Factor (SF)

In MAS 5.0 terminology, Scaling is essentially the multi-chip analog of Normalization. Normalization facilitates the comparison of pairs of arrays by multiplying the intensities on an experimental array so that the trimmed mean of a normalization probe set equals the trimmed mean of the same probe set on the baseline array. In Scaling, all intensities for an array are multiplied by a scaling factor so that the trimmed mean of a scaling probe subset is equal to some specified target intensity. The scaling probe set and target intensity are the same for all arrays.

If scaling is requested, a scaling factor (SF) is calculated as follows and used in the calculation of Signal (see the Expression Estimates section of this website for more information):

$$SF = Sc / \mathit{TrimMean}(\text{signal value}_i, 0.02, 0.98).$$

- Sc = target signal (default Sc = 500)

- $_i$ = probe sets selected for use in the scaling (e.g. known housekeeping genes).
- signal value$_i$ = set of Signals in the array for the scaling probe set
- *TrimMean* = mean of the values excluding the lower and upper 2%

A scale factor of 1.000 is equivalent to no scaling and is the value used in the DFCI Microarray Core. Users are expected to do their own scaling before performing higher-level multi-chip analyses. (Note that in the literature, controlling for brightness of fluorescence intensities across multiple chips is most often referred to as normalization.) Many techniques for normalization exist; see the Normalization section of this website for more details.

Norm Factor (NF)

In MAS 5.0 terminology, Normalization facilitates the comparison of two arrays by controlling for brightness of the chips. A user can specify a normalization factor (NF) for comparing an experiment sample to a baseline sample, and MAS 5.0 will use this in the calculation of Signal (see the Expression Estimates section of this website for more information). If normalization is requested, but the actual NF is not specified, then MAS 5.0 calculates the NF as follows:

NF = *TrimMean*(baseline signal$_i$, 0.02, 0.98) / *TrimMean*(experiment signal$_i$, 0.02, 0.98).

- $_i$ = probe sets selected for use in the normalization calculation (e.g. known housekeeping genes).
- baseline signal$_i$ = set of Signals in the baseline array for the normalization probe set
- experiment signal$_i$ = set of Signals in the experiment array for the normalization probe set
- *TrimMean* = mean of the values excluding the lower and upper 2%

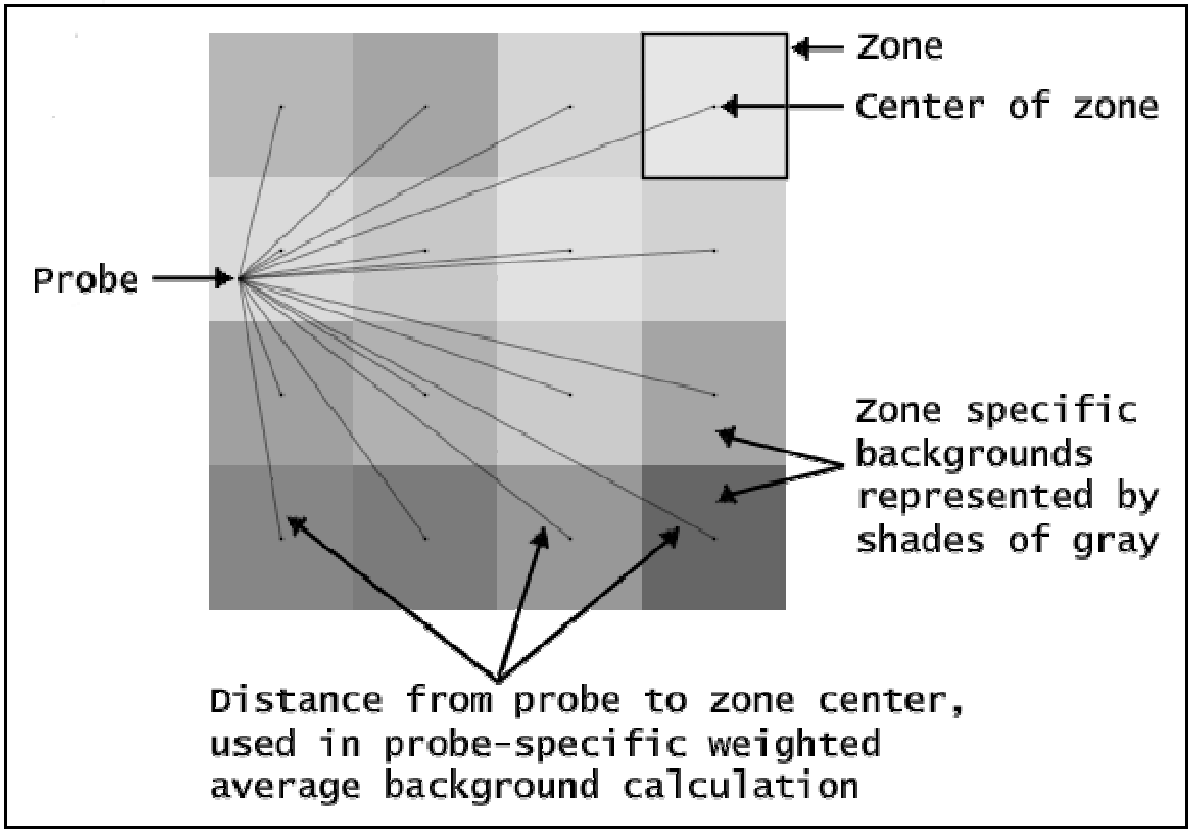
A normalization factor of 1.000 is equivalent to no normalization and is the value used in the DFCI Microarray Core since the absolute analysis algorithm is run separately for each chip. Users are expected to do their own normalization before performing higher-level analyses.

Many techniques for normalization exist; see the Normalization section of this website for more details.

Background

Background is a measure of signal intensity caused by autofluorescence of the array surface and non-specific binding of target/stain molecules. To calculate the background, the array is first divided into 16 zones, and the bottom 2% signal in each zone is used as the background for that zone. A background level for each probe cell is computed by

taking a weighted average of the background for each zone, the weights being determined by the distance of the probe cell to the center of every zone.



The min and max refer to the lowest and highest calculated 2% zone backgrounds. The avg and std are the average and standard deviation of the background signal for the zone centers.

Noise

For the same 16 zones used in the background calculations, the standard deviation of the lowest 2% signal in each zone is calculated. The noise for each probe cell $n(x,y)$ is calculated in a similar fashion as background by taking a weighted average of the zone-specific noise levels, the weights being determined by the distance of the probe cell to the center of every zone.

This noise level $n(x,y)$ is used for calculating background adjusted Signals. For a probe cell at coordinates (x,y), the background adjusted intensity $A(x,y)$ is reported as follows.

$$A(x,y) = \max(I'(x,y)-b(x,y), NoiseFrac*n(x,y))$$

- $I'(x,y) = \max(I(x,y), 0.5)$

- *I(x,y)* = cell intensity at coordinates *(x,y)*
- *b(x,y)* = background intensity at coordinates *(x,y)*
- *NoiseFrac* = selected fraction of global background variation (0.5 by default)

In the .RPT file, min and max refer to the minimum and maximum zone-specific noise levels. Avg and std are the average and standard deviation of the minimum and maximum zone-specific noise levels.

Corner+

Thirty-two probe cells are used for grid alignment; the average intensity of these cells is reported here. This part of the grid is not tiled with probe sequences, so the intensity value is usually approximately equal to background. The value reported here should be much lower than the Corner- value.

Corner-

Thirty-two antisense probe cells are used for grid alignment; the average intensity of these cells is reported here. The value reported here should be much higher than the value reported in Corner+.

Central-

In an antisense probe array (see the Controls entry in this .RPT file), nine probe cells compose a cross at the center of the array and are used for checking grid alignment. The average intensity of those cells is reported here.

Housekeeping Controls

Housekeeping controls can be useful for monitoring the quality of the target sample. The signal and detection call are reported for the probe sets that are specific to the 5', middle, or 3' portion of the transcript. Values >3 or <1/3 for Sig(3'/5') may indicate degradation of the target.

Spike Controls

Spike controls are added to the sample to monitor hybridization, washing, and staining of the chip. Both housekeeping and spike control probes can be used for normalization since we expect that their expression levels should not vary across samples.